AFRL-SR-AR-TR-07-0523

REPORT DOCUMENTATION PAGE

Form Approved OMB No. 0704-0188

PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
07/27/2004	FINAL PROGRESS REPORT	04/01/2005 - 06/30/2007

4. TITLE AND SUBTITLE

APPLICATION OF CORTICAL PROCESSING THEORY TO ACOUSTICAL ANALYSIS

5a. CONTRACT NUMBER	FA9550-05-C-0032
5b. GRANT NUMBER	N/A
5c. PROGRAM ELEMENT NUMBER	N/A

6. AUTHOR(S)

GHITZA, ODED

5d. PROJECT NUMBER	N/A
5e. TASK NUMBER	N/A
5f. WORK UNIT NUMBER	N/A

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

SENSIMETRICS CORPORATION
48 GROVE STREET - SUITE 305
SOMERVILLE, MA 02144-2500

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

AIR FORCE OFFICE OF SCIENTIFIC RESEARCH
975 NORTH RANDOLPH STREET ROOM 3112
ARLINGTON, VA 22203
Dr Willard Larkin/NL

10. SPONSOR/MONITOR'S ACRONYM(S)

AFOSR

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

N/A

12. DISTRIBUTION/AVAILABILITY STATEMENT

N/A

Approved for public release. Distribution is unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

We developed a computational model of diphone perception based on salient properties of peripheral and central auditory processing. The model comprises an efferent-inspired closed-loop model of the auditory periphery (PAM) connected to a template-matching circuit (TMC). Robustness against background noise is provided principally by the PAM, while insensitivity to time-scale variations is provided by the TMC. We demonstrated that for synthetic DRT word-pairs in noise the model is capable of predicting human error patterns along the acoustic-phonetic features. We showed that with a closed-loop PAM, a place/rate model of central processing is sufficient to predict human performance in discriminating speech stimuli in the presence of noise. This result is in contrast to the current notion based upon feed-forward models, which suggests that a temporal (place or non-place) strategy is necessary in order to account for the robust human performance in noise. Towards a generalization to naturally spoken speech we are studying a TMC inspired by principles of cortical neuronal processing, with a gamma rhythm at its core. It falls short when applied to the task of recognizing natural speech, however we demonstrate that it exhibits properties, such as time-scaling insensitivity, consistent with (and desirable for) perception of spoken language.

15. SUBJECT TERMS

Models of auditory periphery; Models of descending auditory pathways; Models of the MOC efferent system; Neural computation approach to modeling post auditory nerve processing; Template matching; Phone discrimination; Phone identification; Diagnostic assessment of speech intelligibility; Front-end for automatic speech recognition.

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			DR. ODED GHITZA
					19b. TELEPHONE NUMBER *(Include area code)* 617-625-0600 X239

Standard Form 298 (Rev. 8/98)
Prescribed by ANSI Std. 239.18

EXECUTIVE SUMMARY

We developed a computational model of diphone perception based on salient properties of peripheral and central auditory processing. The model comprises an efferent-inspired closed-loop model of the auditory periphery (PAM) connected to a template-matching circuit (TMC). Robustness against background noise is provided principally by the signal processing performed by the PAM, while insensitivity to time-scale variations is provided by properties of the TMC. The PAM parameters were determined *in isolation* from the TMC. This was achieved by analyzing confusion patterns generated in a paradigm with a minimal cognitive load (the binary Diagnostic Rhyme Test [DRT], with *synthetic* speech stimuli to restrict phonemic variation). Originally, we intended to test the model by quantifying its ability to predict human performance in perceiving *naturally spoken speech* in the presence of noise, in two separate tasks: (1) diphone discrimination of minimal word-pairs (Voiers' DRT), and (2) phone identification of schwa-CVC tokens. Eventually, the model was evaluated using synthetic speech material.

• **Accomplishments:**

1. For the diphone discrimination task and for the phone identification task we have created a synthetic version of the speech material (DRT word-pairs and schwa-CVC tokens, respectively). Compared to the naturally spoken stimuli, the synthetic stimuli are with restricted phonemic variations, allowing a better PAM-TMC separation.
2. For the diphone discrimination task and for the phone identification task we have completed collecting human-performance data for the synthetic and for the naturally spoken stimuli in noise, using speech-shape noise at three SPL intensities (70, 60 and 50dB) and at three SNRs (10, 5 and 0dB).
3. We demonstrated that, for synthetic DRT word-pairs in noise, the model is capable of predicting both the mean performance and the *patterns of errors* in the human data.
4. We showed that with an efferent-inspired closed-loop model of the cochlea, a place/rate model of central processing is sufficient to predict human performance in discriminating speech stimuli in the presence of noise. This result is in contrast to the current notion based upon feed-forward models, which suggests that a temporal (place or non-place) strategy is necessary in order to account for the robust human performance in noise.
5. To generalize these results to naturally spoken tokens, i.e. tokens that inherently exhibit phonemic variability, we have been studying a template matching circuit that is insensitive to time-scale variations of the input stimuli. We chose to study a template matching circuit (TMC) inspired by principles of cortical neuronal processing, with a gamma rhythm at its core. In its current form the circuit falls short when applied to the task of recognizing naturally spoken speech, however we demonstrate that it exhibits properties, such as time-scaling insensitivity, consistent with (and desirable for) perception of spoken language.
6. For the phone identification task we evaluated four models of frequency-band integration in two experiments on the identification of schwa-CVC syllables. All of the models considered make predictions based on observed confusion matrices. One experiment tested the ability to integrate cues for speech sounds presented in low- and high-frequency bands of speech. The other experiment tested the ability to identify speech sounds presented at different signal to noise ratios. The results of the first experiment are encouraging in terms of the relation of data to model predictions. The results of the second experiment are far less encouraging in terms of the ability of the models to predict dependence of overall scores on signal to noise ratio.

• **Ph.D. dissertation:**

1. Messing, D. (2007). Predicting Confusions and Intelligibility of Noisy Speech. Thesis advisors: L. Braida and O. Ghitza. Department of Electrical Engineering, MIT.

• Publications:

[1] Ghitza, O., Messing, D., Delhorne, L., Braida, L., Bruckert, E., Sondhi, M.M. (2007). Towards predicting consonant confusions of degraded speech. In: Kollmeier, B., Klump, G., Hohmann, V., Langemann, U., Mauermann, M., Uppenkamp, S., Verhey, J. (eds.). *Hearing – From Sensory Processing to Perception,* Berlin: Springer Verlag, in press.

[2] Ghitza, O. (2007). Using auditory feedback and rhythmicity for diphone discrimination of degraded speech. *International Congress of Phonetic Sciences*, August 6-10, Saarbrücken, Germany.

[3] Messing, D., Braida, L. and Ghitza, O. (2007). A non-linear efferent-inspired model of the auditory periphery: signal processing in noise. In preparation. (To be submitted to *Speech Communications*).

[4] Messing, D., Braida, L. and Ghitza, O. (2007). A non-linear efferent-inspired model of the auditory periphery: predicting consonant confusions in noise. In preparation. (To be submitted to *JASA*).

• Personnel:

1. Braida, L. (MIT). Co-PI and advisor to Messing's Ph.D. dissertation.
2. Bruckert, E. (Sensimetrics). Generated a synthetic version of the database (DRT; schwa-CVC).
3. Delhorne, L. (MIT). Designed and administered the behavioral experiments (DRT; schwa-CVC).
4. Ghitza, O. (Sensimetrics). PI and co-advisor to Messing's Ph.D. dissertation.
5. Messing, D. (MIT). Developed the PAM. Designed and conducted all model simulations.
6. Sondhi, M. (Avaya). Consulted on signal processing properties of the TMC.

• Impact:

1. NSF (STTR-Phase I): Exploiting Nervous System Rhythmicity for Spoken Word Recognition. Aimed at recognizing *words* by exploiting the presumed role of nervous-system rhythms in neural computation. With Dr. J. Glass (MIT) as Co-PI.
2. Ph.D. Thesis: Using DRT word-pairs to model a damaged cochlea. Aimed at exploiting the DRT prediction engine developed under this contract to model cochleae of subjects with moderate hearing impairment. Collaboration with Prof. T. Dau (Centre for Applied Hearing Research, Technical University of Denmark).

1. INTRODUCTION

The work described here arose from the need to understand and predict speech confusions caused by acoustic interference and by hearing impairment. Current predictors of speech intelligibility are inadequate for making such predictions (even for normal-hearing listeners). The Articulation Index, and related measures, STI and SII, are geared to predicting speech intelligibility. But such measures only predict average intelligibility, not error patterns, and they make predictions for a limited set of acoustic conditions (linear filtering, additive noise, reverberation).

We aim at predicting consonant confusions made by normally-hearing listeners, listening to degraded speech. Our prediction engine comprises an efferent-inspired peripheral auditory model (PAM) connected to a template-match circuit (TMC). Figure 1 shows a block diagram of the model. We hypothesize that robustness against background noise is provided principally by the signal processing performed by the peripheral circuitry, while insensitivity to time-scale variations is provided by properties of the template-matching circuitry. The extent to which this model is an accurate description of auditory perception is measured within the context of perceiving minimal word pairs (differing in their initial consonant) in the presence of additive, speech-shaped noise. In Section 2 we describe the PAM, a closed-loop model of the auditory periphery that comprises a nonlinear model of the cochlea with efferent-inspired feedback. The PAM parameters were determined in isolation from the TMC. This was achieved by analyzing confusion patterns generated in a paradigm with a minimal cognitive load (Voiers' Diagnostic Rhyme Test [DRT] [17], with *synthetic* speech stimuli to restrict phonemic variation). In Section 3, we describe initial steps towards predicting confusions of *naturally spoken* diphones (i.e. material that exhibits inherent phonemic variability). We describe a TMC inspired by principles of cortical neural processing, with a gamma rhythm at its core (Hopfield, [12]). A desirable property of the circuit is insensitivity to time-scale variations of the input stimuli, a property essential for recognizing phonetic entities that are inherently variable in time and spectrum. In its current form the circuit falls short when applied to the task of recognizing naturally spoken speech, however we demonstrate that it exhibits properties, such as time-scaling insensitivity, consistent with (and desirable for) perception of spoken language. In Section 4 we describe an effort to evaluate four models of frequency-band integration in two experiments on the *identification* of schwa-CVC syllables. All of the models considered make predictions based on observed confusion matrices. One experiment tested the ability to integrate cues for speech sounds presented in low- and high-frequency bands of speech. The other experiment tested the ability to identify speech sounds presented at different signal to noise ratios. The results of the first experiment are encouraging in terms of the relation of data to model predictions. The results of the second experiment are far less encouraging in terms of the ability of the models to predict dependence of overall scores on signal to noise ratio.

2. PERIPHERAL AUDITORY MODEL (PAM)

2.1 Background

A reasonable, axiomatic assumption is that information in the auditory nerve is the only information available to the central nervous system (CNS) about *acoustic* input. While human performance in adverse conditions deteriorates only modestly, *simulated* AN representations of corrupted speech signals - generated by state-of-the-art auditory models - are markedly different from those associated with clean speech signals. For example, for speech in a typically reverberant room, there is only a slight deterioration of intelligibility (albeit with a noticeable degradation in quality) while the acoustic signature of the phonemic features in the simulated AN representations is severely compromised. Is this contrast a result of the incompleteness of current models of auditory processing?

Numerous papers have been published that examine how the response of the cochlea may be processed to provide a relevant representation of the speech signal. Each study utilizes a computational model to simulate either the direct firing activity or another related representation of the cochlear output. The manner in which this information is processed differs among the studies, reflecting differences in the structural properties of the central processor hypothesized by each

study. These structural properties can be cataloged using the following three categories: (1) *place/rate* category, where the central processor possesses explicit knowledge of place (i.e. the fibers' tonotopic place of origin in the cochlear partition) but uses only short-term rate information of the neural firings, over a prescribed time window, (2) *place/temporal* category, where place information is used together with detailed temporal information of local neural responses (i.e. higher-order firing statistics, like the interspike interval statistics), and (3) *non-place/temporal* category, where place information is omitted altogether and the only sources of information are the temporal properties of the global neural response (for an excellent overview of auditory models the reader is referred to [9]). These models of auditory periphery are feed-forward models, based on our understanding of the ascending auditory pathway up through the auditory nerve. A rigorous study of the capabilities of these models to reliably represent speech signals in a variety of acoustic conditions (e.g., different sound intensities, and presence of background noise) reached the widely accepted notion that place/rate models are insufficient, and that (at least) some degree of temporal information is required.

One auditory mechanism that may play a role in the robustness of the auditory periphery in the presence of background noise is the medial olivocochlear (MOC) efferent feedback system. Numerous studies have been published providing detailed morphological and neurophysiological description of the system (e.g. Guinan [10]), as well as psychophysical accounts for its effect on the sensory representation of signals embedded in noise. MOC efferents originate from neurons in the medial superior olivary nucleus (MSO) and terminate directly on outer hair cells (OHC). They have tuning curves similar to, or slightly broader than those of AN fibers (e.g. Guinan [10]), and they project to different places along the cochlear partition in a tonotopic manner. We currently do not have a clear understanding of the functional role of this mechanism. One speculated role, which is of particular interest for the current study, is a dynamic regulation of the cochlear operating point that depends on background acoustic stimulation and which results in robust human performance in perceiving speech in a noisy background. There are a few neurophysiologcal studies consistent with this hypothesis. Using anesthetized cats with noisy acoustic stimuli, Winslow and Sachs showed that by stimulating the MOC nerve bundle electrically, the dynamic range of discharge rate at the AN is partially recovered, [18]. Measuring neural responses of *awake* cats to noisy acoustic stimuli, May and Sachs showed that the dynamic range of discharge rate in cochlear-nucleus neurons is only moderately affected by changes in levels of background noise, [15]. Finally, a few behavioral studies indicate the potential role of the MOC efferent system in perceiving speech in the presence of background noise. Dewson presented evidence that MOC lesions impair monkeys' ability to discriminate between the vowels [i] and [u] in the presence of masking noise, but have no effect on performance in quiet, [4]. More recently, Giraud et al. ([10]) and Zeng et al. ([19]) showed, albeit inconclusively, that the performance of humans with severed MOC feedback results in relatively poor phoneme perception when the speech is presented in a noisy background.

2.2 Efferent-inspired closed-loop model of the auditory periphery

Inspired by this evidence we have developed a closed-loop model of the auditory periphery (i.e. PAM) which uses feedback to regulate the operating point of a model of cochlear mechanics, resulting in an auditory nerve representation less sensitive to changes in environmental conditions. In implementing the PAM we use a bank of overlapping cochlear channels uniformly distributed along the ERB (equivalent rectangular bandwidth) scale, four channels per ERB. Each cochlear channel comprises a nonlinear filter and a generic model of the inner hair cell (IHC) – half-wave rectification followed by low-pass filtering, representing the reduction of neural synchrony with AN fiber characteristic frequency (CF). The dynamic range of the simulated IHC response is restricted to a dynamic-range window (DRW), representing the observed dynamic range at the AN level. The simulated IHC response (representing *instantaneous* firing rate at the AN) is smoothed temporally (temporal time integration over a 10-ms window), resulting in a short-term average-rate representation. (See Fig. 2.)

The cochlear filter is Goldstein's MBPNL model of nonlinear cochlear mechanics, [7]. This model operates in the time domain and changes its gain and bandwidth with changes in the input intensity, in accordance with observed physiological and psychophysical behavior. The model is shown in Fig. 3. The lower path (H_1/H_2) is a compressive nonlinear filter that represents the sensitive, narrowband nonlinearity at the tip of the basilar membrane tuning curves. The upper path (H_3/H_2) is a linear filter that represents the insensitive, broad-band linear tail response of basilar-membrane tuning curves. A parameter GAIN controls the gain of the tip of the basilar membrane tuning curves. To best mimic psychophysical tuning curves of a healthy cochlea in quiet, the tip gain is set to GAIN=40dB [7]. The "iso-input" frequency response of an MBPNL filter at CF of 3400Hz is shown in Fig. 4, upper-left panel.

As for the efferent-inspired part of the model we mimic the effect of the medial olivocochlear efferent path (MOC). Recall that morphologically, MOC neurons project to different places along the cochlear partition in a tonotopic manner, making synapse connections to the outer hair cells and hence affecting the mechanical properties of the cochlea (e.g. increasing basilar membrane stiffness). Therefore, we introduce a frequency-dependent feedback mechanism, which controls the tip-gain of each MBPNL channel, permitting a prescribed intensity level of the sustained noise inside the DRW.

Figure 5 shows – in terms of a spectrogram – simulated IHC responses to diphone ***je*** (as in "jab") in two noise conditions (70 dB SPL / 10 dB SNR and 50 dB SPL / 10 dB SNR), for an open-loop MBPNL-based system (left-hand side) and for the closed-loop system (right-hand side). Due to the nature of the noise-responsive feedback, the closed-loop system produces spectrograms that fluctuate less with changes in noise intensity compared to spectrograms produced by the open-loop system. This property is desirable for stabilizing the performance of template matching under varying noise conditions, as reflected in the quantitative evaluation reported in Section 2.3.

2.3 Quantitative evaluation – isolating PAM from TMC

The evaluation system comprises a PAM followed by a TMC. *Ideally*, to eliminate PAM-TMC interaction, errors due to template matching should be reduced to zero (i.e. ideal template-matching). In *reality* we could only minimize interaction. This was achieved by taking the following three steps: (1) we use the simplest possible psychophysical task in the context of speech perception, namely a binary *discrimination* test. In particular, we use Voiers' DRT ([17]) which presents the subject with a two alternative forced choice between two alternative CVC words that differ in their initial consonants. Such task minimizes the influence of cognitive and memory factors while maintaining the complex acoustic cues that differentiate initial diphones (recall the central role of diphones in speech perception, e.g. Ghitza, [5]); (2) we use the DRT paradigm with synthetic speech stimuli. An acoustic realization of the DRT word-pairs was synthesized so that the target values for the formants of the vowel in a word-pair are identical, restricting stimulus differences to the initial diphones; and (3) we use a "frozen speech" methodology (e.g. Hant and Alwan, [11]), namely, the same acoustic speech token is being used for training and for testing, so that testing tokens differs from training tokens only by the acoustic distortion.

These three steps presumably result in a reduction in the number of errors induced by the template matching. Recall that a template-match operation comprises measuring the distance of the unknown token to the templates, and labeling the unknown token as the template with the smaller distance. Hence, template matching is defined by the distance measure and the choice of templates. As a distance measure we use the minimum mean squares error. This is an effective choice here because: (1) by using synthetic speech stimuli, the identical target values of the vowel formants for the two words results in zero error in time-frequency cells associated with the final diphone, and (2) by using frozen-speech stimuli, a distortion in a given time-frequency cell is generated *locally* (by noise component within the range of the cell) and is independent of noise at other cells. Thus, with such constraints it was reasonable to use a template-matching operation with a minimum mean squares error as the distance measure, allowing us to focus on errors attributed to the PAM alone.

Formal DRT sessions using human subjects have been conducted using the synthetic stimuli in quiet and in additive, speech-shaped noise at three levels (50, 60 and 70 dB SPL) and at three SNRs (0, 5 and 10 dB). The psychophysical experiments are described in detail in Appendix A. Fig. 6 shows the errors produced by a DRT mimic with open-loop and closed-loop PAMs, compared to those made by human listeners. Figure 6.a shows performance averaged over all SPL and SNR conditions. (Figures 6.b and 6.c detail the performance as per SPL×SNR condition.) Signal conditions were the same as those used to collect the human data. Templates were created for the 60 dB SPL / 5 dB SNR condition. The abscissa marks the Jakobsonian dimensions: Voicing, Nasality, Sustention, Sibilation, Graveness and Compactness (denoted VC, NS, ST, SB, GV and CM, respectively). The "+" sign stands for an attribute being present and the "-" sign for an attribute being absent. Bars show the *difference* between the average machine and human scores. The red "boxes" indicate plus and minus one standard deviation of the human data. Gray bars indicate that the difference is greater than one standard deviation of the human data. Scores with the open-loop PAM are worse than those of the human scores. Scores with the closed-loop PAM are similar to human scores except for VC- and ST-. Two points are noteworthy. First, when a sever mismatch occurs, closed-loop scores are *superior* to human scores while open-loop scores are worse. Hence, improving the open-loop system will require the exploitation of information beyond short-term rate (i.e. *temporal*). Second, although we predicted human performance in a binary task, parameters of the model were tuned to match errors between minimal pairs, *jointly* along *all* Jakobsonian dimensions. Hence we believe that the spectro-temporal patterns generated by the resulting closed-loop PAM are an adequate description of the sensory representation of degraded speech.

3. THE TEMPLATE MATCHING CIRCUIT (TMC)

In developing the PAM (Section 2) we used synthetic speech stimuli, with restricted phonemic variation, hence permitting the use of a minimum mean squares error as the distance measure for template matching. In this section we consider naturally spoken speech stimuli, seeking a perceptually relevant distortion measure between speech tokens that exhibit phonemic variability.

3.1 Why use models of neural computation?

In some sense speech decoding can be conceptualized as a search process, in which the search engine performs a template-matching operation comprised of two separate, but related steps. The first measures the *distance* between the current input (e.g. a syllable) and (stored) templates. The second associates the input with the best-matching template. In this sense, template matching is defined by the *choice* of templates as well as a *distance metric*. To develop algorithms capable of emulating human performance we first need to create accurate, detailed models for both stages of the search process. An explicit, *analytical* expression is difficult to derive for such models. Instead, we seek to emulate neural computation principles that are general in nature and shared across sensory (e.g. auditory, visual, olfactory) and motor modalities. We suggest that a template-matching operation based on a *plausible* model of pertinent neural computation may implicitly incorporate characteristics essential for both the templates and the distance metric. Next we describe a specific template-matching circuit inspired by principles of cortical neural processing, with a gamma rhythm at the core (Hopfield, [12]). This oscillation feeds into all input neurons, serving as a synchronizing pacemaker.

3.2 Model description

A block diagram of the TMC is shown in Figure 7. It comprises three stages: (1) a front-end, (2) a layer of "Integrate and Fire" (IAF) neurons (Layer-I neurons) and (3) a layer of coincidence-detector neurons (Layer-II neurons). The front-end is a filter bank with 26 critical-band filters spanning the tonotopic range of the speech spectrum (0.1 – 8 kHz). Each neuron in Layer-I is characterized by the equation:

$$du(t)/dt + u(t)/RC = i(t)/C - Vrest/RC \qquad (1)$$

where *i(t)* is the input current, *u(t)* the output voltage, *Vrest* is the resting potential and *RC* is the time constant of the circuit. Once *u(t)* reaches a threshold value the neuron fires and *u(t)* is shunted to zero. The parameters of all Layer-I neurons are identical except for the threshold-of-firing. All Layer-I neurons are driven by a single global sub-threshold oscillatory current $A_\gamma cos\ \omega_\gamma t$. In terms of Eq. (1), the input current to the *n*-th IAF cell is:

$$i_n(t) = x_n(t) + A_\gamma cos\ \omega_\gamma t \qquad (2)$$

where $x_n(t)$ is the output of the *n*-th cochlear channel. In our realization, RC = 20 ms and the frequency of the gamma oscillator is 25 Hz. Each channel drives 100 Layer-I neurons, which differ only in their firing threshold. In our realization, the number of Layer-I neurons is M = 26 × 100 = 2600. The final stage comprises N = 6000 Layer-II coincidence neurons. All Layer-II neurons are driven by K=6 randomly selected "patches" of Layer-I neurons. Each patch is composed of L=10 Layer-I neurons with successive thresholds – all driven by the same frequency channel.

The computational principle realized by the circuit is as follows. A given Layer-II neuron fires at time t_0 if and only if all K Layer-I patches fire simultaneously at time t_0. Moreover, a patch of Layer-I neurons fires at time t_0 only if the time evolution of the corresponding frequency channel prior to that time drives one of the L neurons in the patch to its threshold precisely at time t_0. Hence, each Layer-II neuron is "tuned" to a particular time-frequency template expressed in terms of the time evolution of K frequency channels. The same Layer-II neuron will also fire, albeit in a delayed time, if the output signal of all K channels is scaled by the same factor (this is so because all corresponding Layer-I neurons reach threshold with a similar time delay).

This TMC has some interesting properties germane to speech processing and decoding. As illustrated in Section 3.3.2 syllable-initial diphones (i.e. consonant-vowel, CV syllables) may be identified more accurately than their syllable-final (i.e. coda, VC) counterparts. This property is consistent with both linguistic perception and with statistical analyses of conversational corpora where spectro-temporal variability of coda consonants is far greater than their consonantal counterparts in syllable onsets (Greenberg, [8]). Moreover, variation in speaking rate has relatively little impact on its performance since the TMC is insensitive to time-scale variation (consistent with Hopfield's original formulation). Such time-scale insensitivity (e.g. to variation in speaking rate) is essential for recognizing phonetic entities that are inherently variable in time and spectrum. These are the sort of properties that characterize human speech comprehension and which could prove useful for many technical applications in speech recognition, synthesis and auditory prostheses.

3.3 Illustrative Examples

3.3.1 TMC response in a simple syllable-discrimination task

Figure 8 illustrates the behavior of the TMC in a simple syllable-discrimination task. Assume that we have identified 40 Layer-II neurons that are most sensitive to the time-frequency signature of the initial diphone of the word "daunt." Similarly, we have identified 140 neurons for the word "taunt." We term these sets of cells "State-1" and "State-2" neurons, respectively. The two upper-left panels show a spectrographic display of the front-end in response to the first 350 ms of two different realizations of the word daunt spoken by a single speaker (note the phonetic variability). Below each spectrogram is a time-histogram of the number of state neurons responding to the corresponding stimulus (shown is the pertinent fraction out of 40). The lower-right four panels show the analogous display for the response of State-2 neurons to the word taunt. The lower-left (and the upper-right) panels show the response of the neurons to the other word. The response to stimuli matched to the state neurons peaks at a time-instance associated with the end-time of the initial diphone. For stimuli of the other token there is a relatively small response.

3.3.2 Response of a single frequency channel to a DC input

We used an array of 100 Layer-I neurons as specified in Equation (1). The parameters of all neurons were identical except for their firing thresold. The 100 threshold levels were equally distributed over a pre-specified range. All Layer-I neurons were driven by a single, sub-threshold,

25-Hz oscillatory current. The values of all parameters were normalized with respect to the resting potential of the neuron [*Vrest* of Eq. (1)]. These parameters include (i) the dynamic range of the input signal [*i(t)* in Eq. (1)], (ii) the amplitude of the gamma oscillator [A_γ of Eq. (1)], (iii) the threshold-of-firing pre-specified range. For every measurement point we recorded how the firing patterns of the 100 neurons change as a function of the input current. Figure 9 illustrates the kinds of recording collected (see the figure legend for details).

3.3.3 Response of an array of 100 Layer-I neurons to a saw-tooth

Figure 4 illustrates the response of an array of 100 Layer-I neurons to an asymmetric saw-tooth input current, over a range of "symmetry" coefficients. These coefficients pertain to how fast (or slow) the current rises in time, which has a significant impact on the temporal distribution of neuronal spikes evoked by supra-threshold signals. Neurons are the same as those described in Section 3.3.2. The rationale for using a saw-tooth input is as follows. Recall that the input current to a Layer-I neuron is a narrow-band signal (e.g. the output of an auditory channel). For a consonant-vowel syllable the temporal evolution of the energy at the output of a particular channel resembles a saw-tooth function (as a spectral peak moves through the frequency band). See Figure 10 for additional details.

The asymmetrical response shown in Fig. 10 highlights an important corollary of this modeling approach. Sharply rising input waveforms are associated with a more precise correspondence between input signal and neural spikes. For rapidly changing spectra (as is common in speech) this means that there's a tighter temporal correspondence between the dynamic aspects of the signal and its representation in the cortex.

4. PHONE IDENTIFICATION

We evaluated several models of frequency-band integration in two experiments on the identification of schwa-CVC syllables. All of the models considered make predictions based on observed confusion matrices. One experiment tested the ability to integrate cues for speech sounds presented in low- and high-frequency bands of speech. The other experiment tested the ability to identify speech sounds presented at different signal to noise ratios.

4.1 Materials

Both natural and synthetic CVC syllables were used. The natural CVC materials (Lippmann *et al.*, [14]) were recorded by one male and one female talker. The vowels were the three cardinal vowels plus their unstressed cognates. There were 12 consonants, /p, t, k, b, d, g, s, sh, v, h, z, voiced th, unvoiced th/. Materials were low-pass filtered at 9 kHz then converted to 12 bit digital samples at a sample rate of 20 kHz. The mean durations of the syllables spoken by each of the two talkers were 634 and 574 ms. Synthetic analogs of each of the items were produced by HLSyn, a modification of the Klatt synthesizer that was developed by Sensimetrics Corporation. There were 840 items in each of the natural and synthetic sets of the test syllables.

The initial and final consonants for a CVC token were independently drawn with probability 1/12 from the set of 12 consonants, allowing for duplications of CVCs (one male and one female token) and omissions. Each CVC stimulus was presented a total of roughly 70 times to each listener under each condition. Thus there were a total of 840 total stimulus presentations under each condition.

For presentation, the CVC materials were downsampled to a bandwidth of 5000 Hz. All filtering was performed by linear-phase FIR filters with transition region widths were 50 Hz, with out-of-band attenuations of 80 dB. All stimuli were presented binaurally at 60 dB SPL.

4.2 Human performance (see also Appendix B.)

Listeners

The listeners in all experiments were young adults with clinically normal hearing (thresholds better than 20 dB HL at audiometric test frequencies from 200 to 8000 Hz). In both Experiments I and II

there were four listeners (L1, L2, L4 were female and L3 male). Two of the listeners responded to the natural stimuli and the other two to the synthetic stimuli.

Experiments
Experiment I was intended to investigate the ability of listeners to integrate cues for consonants from two, relatively wide, bands of speech. In Experiment I, materials were filtered into bands 0-2100 Hz. (L) and 2100-4500 Hz. (H) and their sum (B). All materials were presented at 60 dB SPL at a wideband SNR of +5 dB.

Experiment II was intended to investigate the ability of listeners to identify consonants as a function of signal to noise ratio. In Experiment II, materials were filtered into 0-4500 HZ (B) and presented at 60 dB SPL at SNRs of +11, +5, and –1 dB.

4.3 Data Analysis
The data from the two experiments were analyzed by constructing confusion matrices for each listener and each filtering and presentation condition, averaged over testing run and vowel. Data were analyzed separately for initial and final consonants and vowels, by averaging over the presentations of each of these groups of speech sounds. These matrices were then analyzed to determine an overall percentage of correct responses.

4.4 Integration Models
All of the models considered make predictions based on observed confusion matrices. Let $Pk(Rj|Si)$ denote the probability that response Rj is made when the stimulus Si is presented via band K and $Pk(R|S)$ denote the confusion matrix for band K. In all cases we use the relative frequency of observing response Rj when stimulus Si is presented to estimate $P(Rj|Si)$. Given the confusion matrices for bands K and L, the models make predictions for the confusion matrix when these bands are presented simultaneously, $Pkl(R|S)$. We estimate the predicted probability of a correct response in the combined-band condition as Ckl.

We compared predictions of probability of a correct response in the combined-band condition using four integration models:

1) According to the Multiplicative Probability of Error model (MPE - Fletcher) a response error is made in the combined-band condition if and only if an error is made in each of the component band conditions. Rather than apply this prediction to overall scores, we applied it separately to each of the stimuli to be identified.
2) According to the Post-Labelling Integration Model (PostL) the listener is assumed to make separate identification judgments about the stimulus based on the cues available in each band, and to combine these judgments to determine the response to the multiband stimulus.
3) According to the Fuzzy Logic Model of Perception (FLMP – Cohen and Massaro, [3]), the response to each stimulus is determined, in a probabilistic fashion, by the "feature value" of that stimulus for each of the possible responses. In single-band presentations the feature value is estimated as the conditional probability, $Pk(Rj|Si)$. In multiband conditions, the feature value is assumed to be proportional to the product of the feature values for the corresponding unimodal conditions (Cohen and Massaro, [3]).
4) According to the Pre-Labelling Integration Model (PreL - Braida, [2]) single-band sensory data are assumed to be represented in continuous valued cues that are combined optimally before labels are assigned. The statistical properties of the cues are inferred from single-band confusion matrices using a type of multidimensional scaling (Braida, [1]). The predictions of the Pre-Labeling Model were made in accordance with the findings of Ronan *et al.* ([16]). Consonant confusion matrices were scaled in four dimensions, the vowels in three. Predictions were made with the response centers half-way between the old response centers and the stimulus centers.

These models can be extended to predict the effects of changing signal-to-noise ratios (SNRs) on identification performance. When stimuli are observed twice in statistically independent noise, the

effect of the noise can be reduced if the observations are combined appropriately. Under optimum conditions, the effective SNR is improved by the square root of two (3 dB) for each doubling of the number of observations. This can be used as the basis of model predictions. Rather than combine two different bands, the models can be applied to combining a band with itself for each 3 dB increase in SNR, or four times for a 6 dB increase in SNR.

4.5 Results

Table I presents the results of, and predictions for, Experiment I. Listeners L1 and L2 were presented with natural (Nat) stimuli, L3 and L4 with synthetic (Syn) stimuli. Observed scores for the low (Obs-L), high (Obs-H), and both (Obs-B) bands are presented for final (F) and initial (I) consonants and vowels (V). Also presented are the predicted scores for the Fuzzy Logic Model of Perception (FLMP), the Multiplicative Probability of Error (MPE), the Post Labeling model (PostL) and the Pre Labeling model (PreL). Finally the average error (Err. (Obs-Prd)) and root-mean-square error (Rms Err.) are shown, both separately for natural and synthetic stimuli and overall (OA Err. and OA Rms Err.).

Overall the results for Experiment I indicate that three of the models (FLMP, MPE, and PostL) tended to under-predict while the PreL model over predicted results. In all cases but PostL the average bias was less than two percentage points. Overall, the smallest rms error was produced by the PreL (3.7 points) followed by the FLMP (4.6), MPE (5.4) and PostL (6.7) points. Similar trends were seen for the subset of natural stimuli, although for the synthetic subset, the rms error was considerably higher for the PreL (4.2 points) and lower for the FLMP (2.2). These results are similar to those of Ronan *et al.* ([16]) for natural CVCs presented in quiet. In particular, for the naturally produced speech sounds, for the low frequency band scores for final consonants were lower on average than scores for initial consonants, whereas there was very little difference between these scores for the high frequency band.

Table II presents the results of, and predictions for Experiment II. In this case predictions of identification scores are made from a base condition (Base, +5 or –1 dB SNR) and to an observed condition (Obs, +11 or +5 dB SNR). We evaluated predictions for the same four integration models (FLMP, MPE, PostL, and PreL) as for Experiment I. Overall the results for Experiment II indicate that all of the models tended to over-predict observed scores between 5.3 (FLMP) and 9.5 (MPE) points. This tendency for over prediction was seen for both natural and synthetic stimuli and for both prediction types (+11/5 and 5/-1 dB SNR). Overall, rms errors were larger than those observed in Experiment I. The smallest rms error was produced by the FLMP (7.2 points) followed by the PreL (9.1), PostL (9.2) and MPE (11.0) points. The trends in rms error were different for the natural and synthetic subsets of stimuli. For the natural subset, the rms error for the PreL was lower than for the FLMP, whereas for the synthetic subset, the reverse was true.

4.6 Summary

The results of Experiment I support the encouraging findings of Ronan *et al.* ([16]), both in terms of data and the relation of data to model predictions. The results of Experiment II are far less encouraging in terms of the ability of the models to predict dependence of overall scores on signal to noise ratio.

The experimental results have only begun to be examined. Before publication, we plan to examine the ability of the models to predict the pattern of correct responses across phonemes in detail. We also plan to develop computational models that make predictions based on the actual acoustic stimuli.

Table I. Results of and Predictions for Experiment I.

	Nat/Syn	F/I/V	Lis	Obs-L	Obs-H	Obs-B	FLMP	MPE	PostL	PreL
	Nat	F	L1	47.5	50.4	80.6	80.7	74.1	68.6	77.1
	Nat	F	L2	42.9	52.6	80.6	74.5	72.5	71.6	75.4
	Nat	I	L1	66.6	53.3	88.4	88.5	84.8	81.0	87.8
	Nat	I	L2	51.8	51.0	84.8	71.2	75.1	72.7	81.2
	Nat	V	L1	95.3	36.7	96.7	98.3	97.8	95.4	97.0
	Nat	V	L2	83.8	39.1	85.9	86.6	92.6	84.6	89.1
Err. (Obs-Prd)							2.9	3.3	7.2	1.6
Rms Err.							6.1	6.6	8.4	3.2
	Syn	F	L3	22.0	32.8	45.0	43.6	48.9	45.8	48.2
	Syn	F	L4	33.6	41.4	63.7	63.0	63.2	55.5	64.6
	Syn	I	L3	49.2	45.0	74.9	71.8	72.4	72.0	78.1
	Syn	I	L4	65.2	54.5	84.4	88.5	85.3	83.1	90.2
	Syn	V	L3	92.3	49.4	89.2	88.2	97.6	94.8	96.2
	Syn	V	L4	96.7	64.7	99.5	99.5	99.1	98.1	99.5
Err. (Obs-Prd)							0.3	-1.6	1.2	-3.4
Rms Err.							2.2	4.0	4.3	4.2
OA Err.							1.6	0.9	4.2	-0.9
OA Rms Err.							4.6	5.4	6.7	3.7

Table II. Results of and Predictions for Experiment II.

Prediction	Nat/Syn	F/I/V	Lis	Base	Obs	FLMP	MPE	PostL	PreL
+11 from +6	Nat	F	L1	80.6	86.7	97.3	98.8	94.8	89.3
+11 from +6	Nat	F	L2	80.6	88.0	91.7	91.6	95.1	92.2
+11 from +6	Nat	I	L1	88.4	92.0	98.4	99.6	96.1	98.1
+11 from +6	Nat	I	L2	84.8	89.5	91.6	92.1	96.8	91.6
+11 from +6	Nat	V	L1	96.7	97.3	100.0	100.0	99.7	99.5
+11 from +6	Nat	V	L2	85.9	90.3	89.8	98.6	92.1	97.3
Err. (Obs-Prd)						-4.2	-6.1	-5.1	-4.0
Rms Err.						6.0	6.8	6.2	3.8
+11 from +6	Syn	F	L3	45.0	62.1	58.2	77.8	76.6	76.1
+11 from +6	Syn	F	L4	63.7	77.4	85.3	91.5	91.6	89.3
+11 from +6	Syn	I	L3	74.9	82.5	90.9	93.9	93.9	91.0
+11 from +6	Syn	I	L4	84.4	86.9	91.6	97.0	93.5	79.3
+11 from +6	Syn	V	L3	89.2	93.4	100.0	100.0	99.0	99.8
+11 from +6	Syn	V	L4	99.5	99.2	100.0	100.0	100.0	100.0
Err. (Obs-Prd)						-4.1	-9.8	-8.8	-5.7
Rms Err.						6.5	12.0	11.1	10.1
+6 from -1	Nat	F	L1	75.3	80.6	95.7	97.6	90.0	93.6
+6 from -1	Nat	F	L2	73.0	80.6	91.4	92.3	90.4	91.3
+6 from -1	Nat	I	L1	84.3	88.4	99.6	99.5	95.3	96.1
+6 from -1	Nat	I	L2	79.3	84.8	91.1	91.4	92.6	91.0
+6 from -1	Nat	V	L1	96.0	96.7	100.0	100.0	99.7	99.9
+6 from -1	Nat	V	L2	87.8	85.9	94.2	99.2	94.5	92.8
Err. (Obs-Prd)						-9.2	-10.5	-7.6	-8.0
Rms Err.						10.2	11.0	7.8	8.9
+6 from -1	Syn	F	L3	36.9	45.0	43.9	67.8	61.0	53.9
+6 from -1	Syn	F	L4	46.1	63.7	62.7	73.5	75.0	79.3
+6 from -1	Syn	I	L3	66.2	74.9	87.0	92.2	91.2	87.4
+6 from -1	Syn	I	L4	74.4	84.4	86.3	92.9	93.5	91.2
+6 from -1	Syn	V	L3	90.8	89.2	100.0	100.0	99.5	99.8
+6 from -1	Syn	V	L4	98.9	99.5	100.0	100.0	100.0	83.3
Err. (Obs-Prd)						-3.9	-11.6	-10.6	-6.4
Rms Err.						7.3	14.9	12.9	11.3
OA Err.						-5.3	-9.5	-8.0	-6.0
OA Rms Err.						7.2	11.0	9.2	9.1

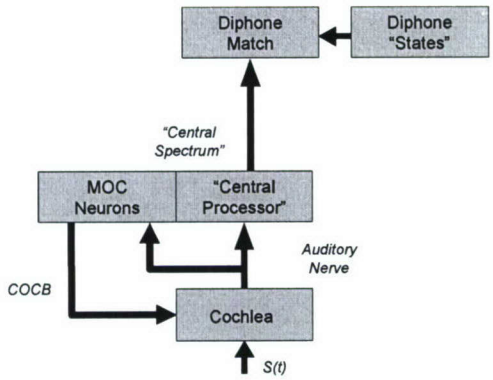

Figure 1: A block diagram of the prediction engine

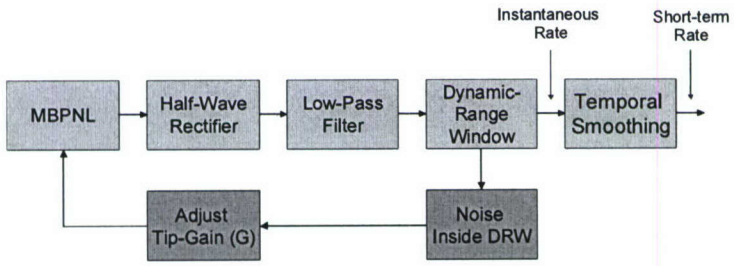

Figure 2: A block diagram of one cochlear channel. The central processor uses place/rate strategy

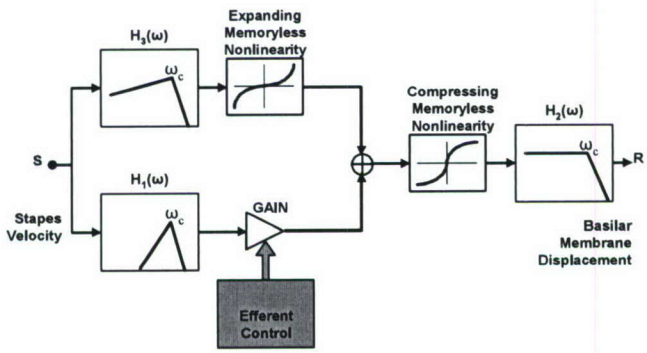

Figure 3: Goldstein's multi bandpass nonlinearity model, MBPNL, [7].

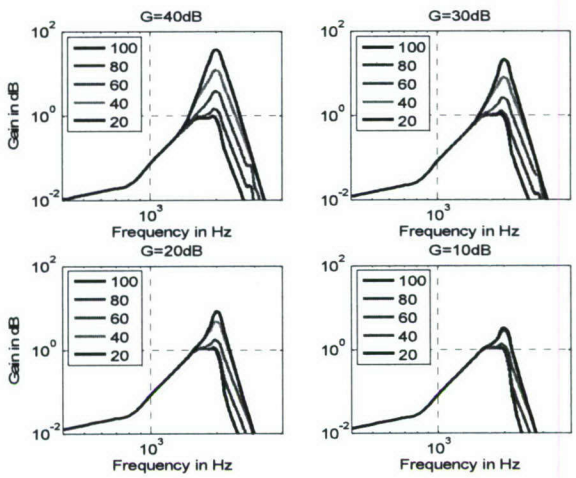

Figure 4: "Iso-input" frequency response, CF=3400Hz. Inside box are input levels in dB SPL.

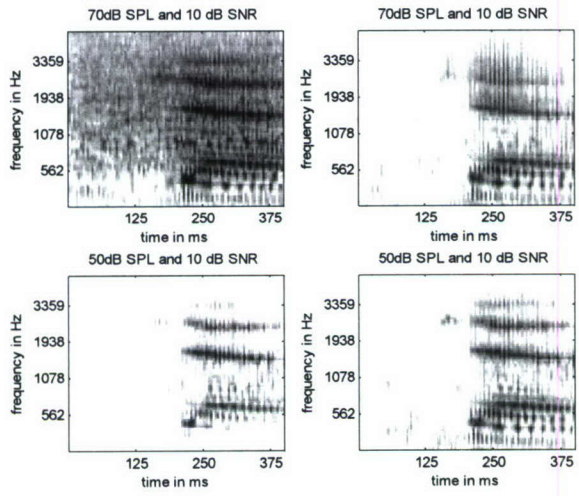

Figure 5: Simulated IHC response for open-loop (left) and closed-loop PAM (right).

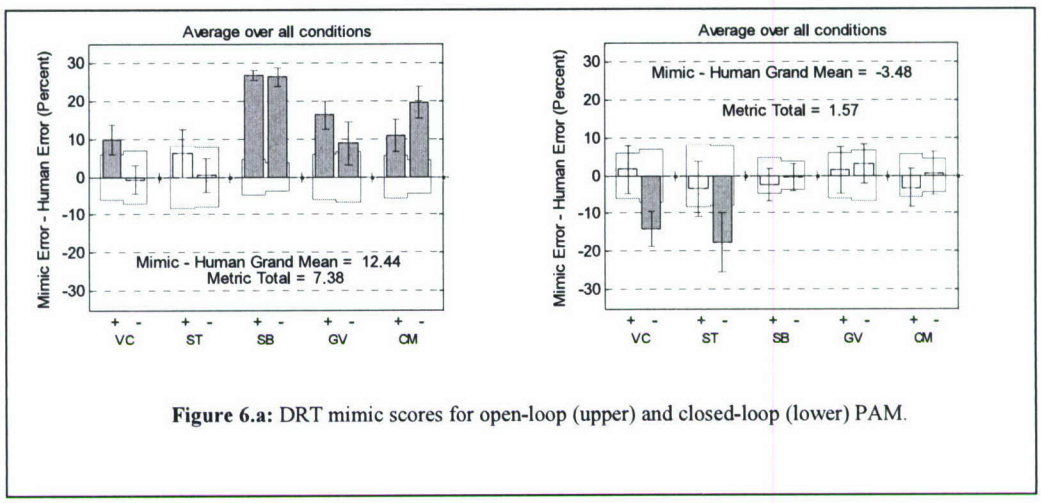

Figure 6.a: DRT mimic scores for open-loop (upper) and closed-loop (lower) PAM.

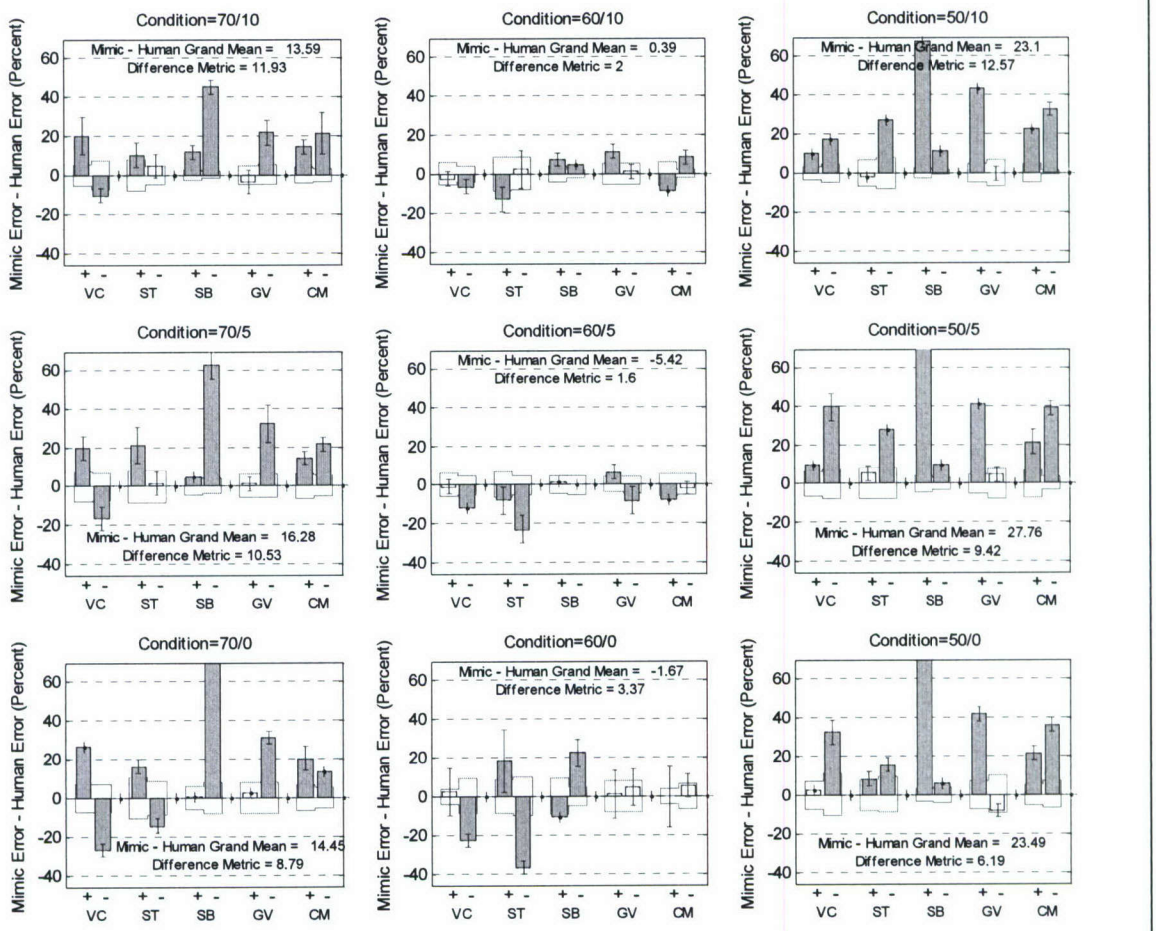

Figure 6.b: 3 SPL x 3 SNR conditions, detailed results for open-loop PAM. Overall, scores per acoustic dimension were different from human and adversely contributed to the within-1-std metric score.

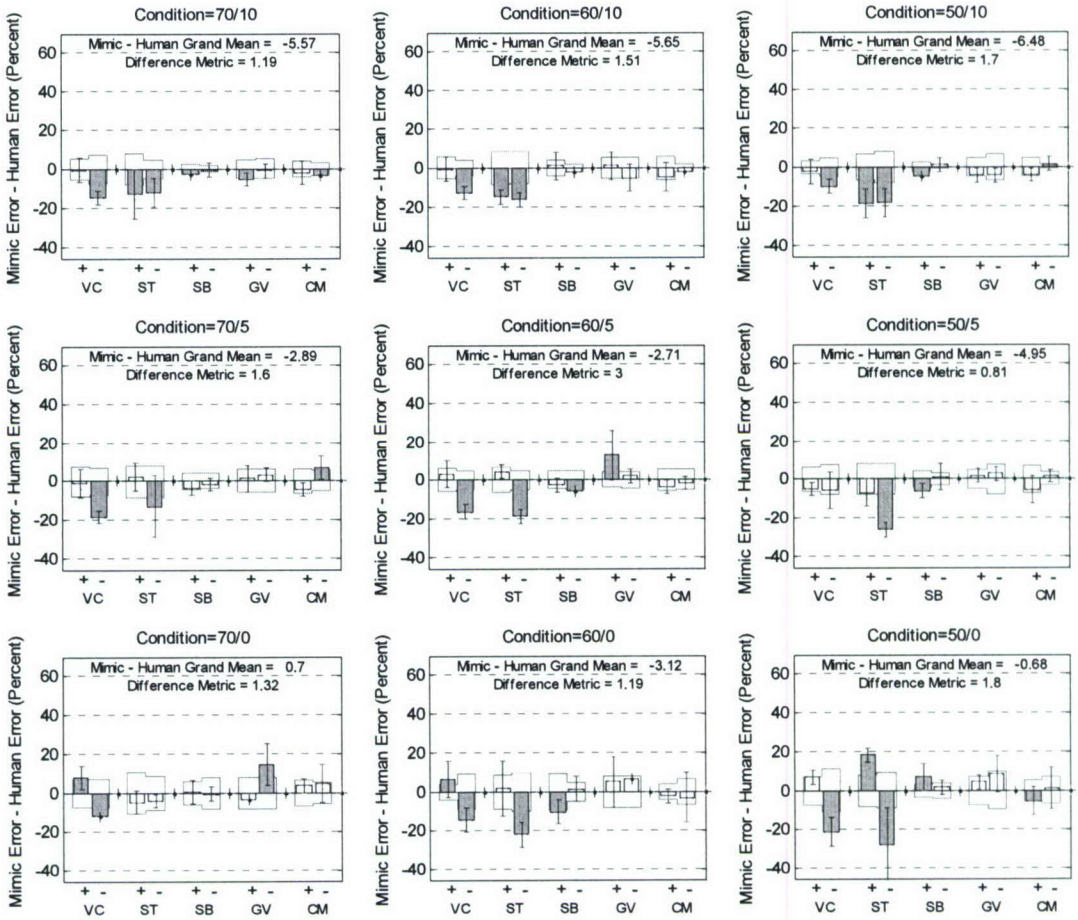

Figure 6.c: 3 SPL x 3 SNR conditions, detailed results for closed-loop PAM. Overall, scores per acoustic dimension were different from human and adversely contributed to the within-1-std metric score.

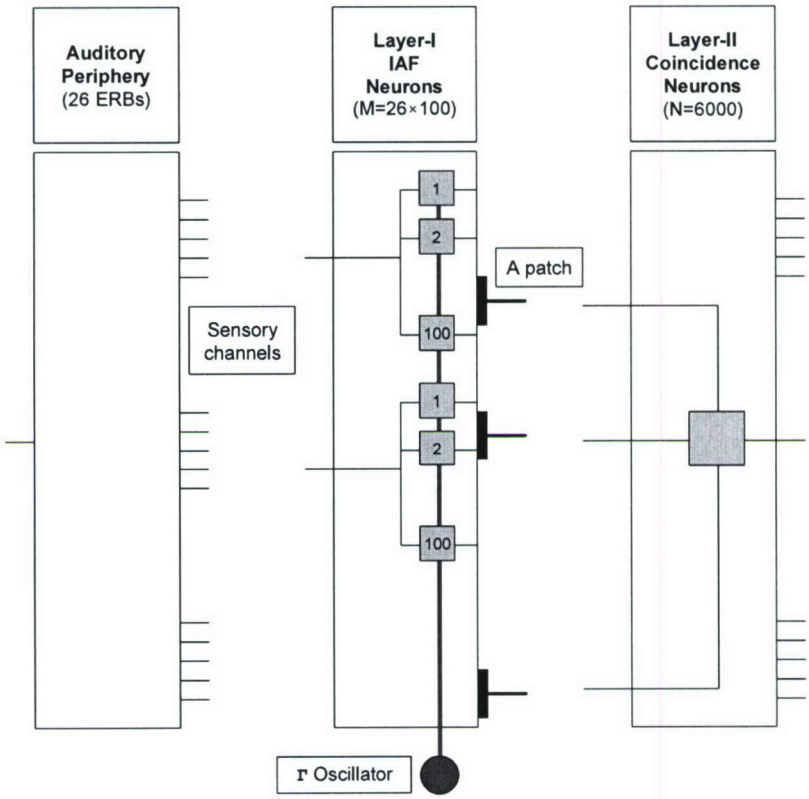

Figure 7: A block diagram of the TMC. The front-end is a filter bank with 26 critical-band channels spanning the range of the speech spectrum. Each channel drives 100 Layer-I integrate-and-fire neurons. The parameters of all Layer-I neurons are identical except for the threshold-of-firing. All Layer-I neurons are driven by a single, global, sub-threshold oscillatory current with a frequency in the gamma range. Each Layer-II coincidence neuron is driven by six randomly selected "patches" of Layer-I neurons.

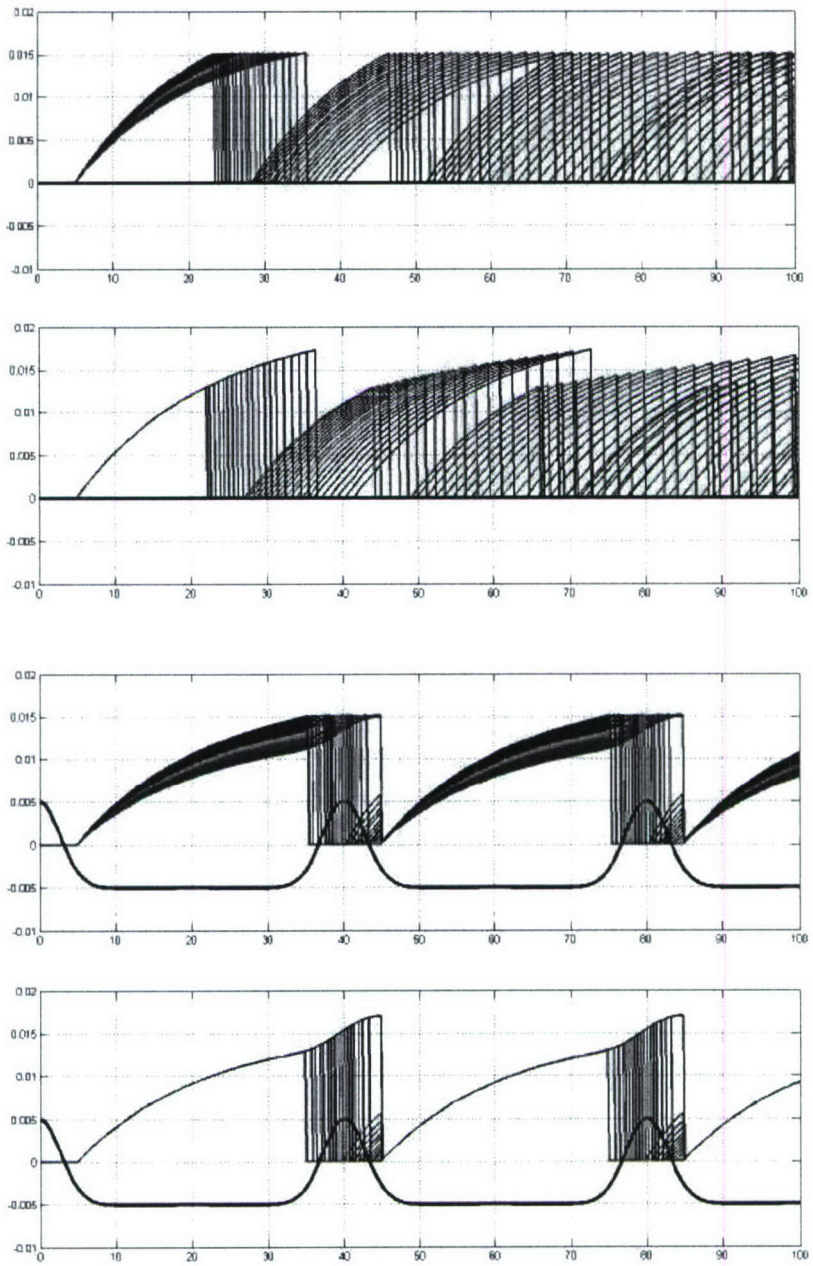

Figure 8: (a) Response of a single neuron to a D.C. input *without* a gamma oscillator. The abscissa corresponds to time in milliseconds. Ordinate is the action-potential of the neuron. Different traces are for different input values. When the action potential reaches threshold the neuron fires and the action-potential resets to zero. Refractory time is 5ms. Time constant is 20ms. (b) The response of an array of neurons (differ only in threshold-of-firing), without a gamma oscillator, to a fixed D.C. input level. Different traces are for different neurons (c) Same as (a) with the inclusion of a gamma oscillator of frequency 25Hz. Note the synchronizing effect of the gamma oscillator. Note also that the gamma oscillator is not a pure sinusoid. (d) Same as (b), but with the inclusion of a gamma oscillator.

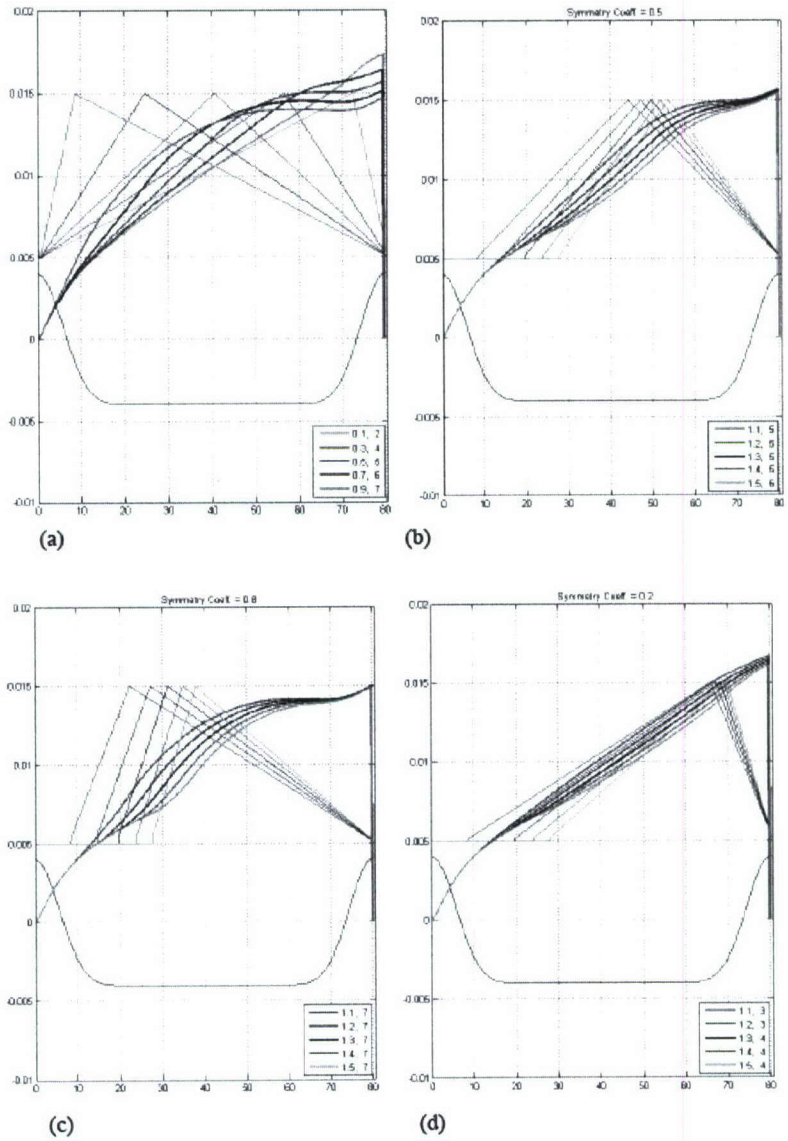

Figure 9: (a) Response of an array of 100 Layer-I neurons to a saw-tooth input currents. The neurons are labeled as 10 successive patches, with 10 successive neurons per patch. All the neurons were driven by one gamma oscillatory current with frequency of 25Hz. Abscise is time in milliseconds. Bottom trace shows one cycle of the gamma oscillation. Five input saw-tooth currents are shown, with different asymmetry (symmetry coefficients are shown in the left column of the legend box). Coupled with each input current is the action potential of all neurons which fire at the time instant corresponding to the peak of the gamma oscillation. The patch number of these firing neurons is shown in the right column of the legend box. Note that each saw-tooth current is mapped onto a different patch. (b) Response to a time-compressed symmetrical saw-tooth (symmetry coefficient of 0.5). The columns of the legend box show the time scaling factor (left) and the patch number of the firing neurons (right), respectively. Note that input currents time-compressed to up to 40% are all mapped onto the same patch. (c) Same as in (b) for a saw-tooth current with symmetry coefficient of 0.8. (d) Same as in (b) for a saw-tooth current with symmetry coefficient of 0.3. Note that a fast rising saw-tooth (panel (c)) is represented with a better precision compared to a slow rising saw-tooth (panel (d)).

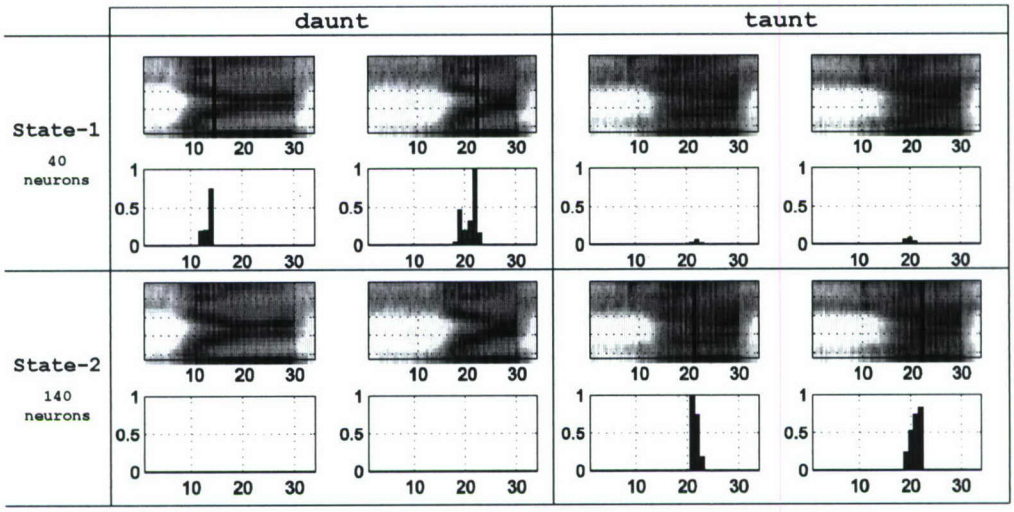

Figure 10: Performance of the TMC in the DRT task. State-1 represents 40 Layer-II neurons most sensitive to the initial diphone of the word "daunt." Analogously, State-2 represents 140 neurons for the word "taunt." The two upper-left panels show a spectrographic display of the front-end in response to the first 350 ms of two different realizations of the word "daunt". Below each spectrogram is a time-histogram of the number of State-1 neurons responding to the corresponding stimulus (shown is the pertinent fraction out of 40). The lower-right four panels show the analogous display for the response of State-2 neurons to the word "taunt". The lower-left (and the upper-right) panels show the response of the neurons to the other word. Note the strong response to stimuli of matched tokens (and weak response to opposite tokens).

APPENDIX A. DRT – HUMAN PERFORMANCE

A.1 Voiers' DRT

The DRT (Diagnostic Rhyme Test) version of Voiers ([17]) is a way of measuring the intelligibility of processed speech and has been used extensively in evaluating speech coders. From an acoustic point of view, Voiers' DRT database covers initial dyads of spoken CVCs. The database consists of 96 pairs of confusable words spoken in isolation. Words in a pair differ only in their initial consonants. The dyads are equally distributed among 6 acoustic-phonetic distinctive features and among 8 vowels (hence 2 word-pairs per [quadrant×feature] cell). The feature classification (outlined in Table A.1) follows the binary system suggested by Jakobson, Fant and Halle (Jakobson *et al.*, [13]), and the vowels are [ee] and [it] (High-Front), [eh] and [at] (High-Back), [oo] and [oh] (Low-Front) and [aw] and [ah] (Low-Back). In our version of the DRT the vowels are collapsed into 4 quadrants (High-Front, High-Back, Low-Front, Low-Back), hence 4 word-pairs per a [quadrant×feature] cell.

The psychophysical procedure is carefully controlled to assure a task with minimum cognitive load. The listeners are well trained and are very familiar with the database, including the voice quality of the individual speakers. The experiment uses a one-interval two-alternative forced-choice paradigm. First, the subject is presented visually with a pair of rhymed words. Then, one word of the pair (selected at random) is presented aurally and the subject is required to indicate which of the two words was played. This procedure is repeated until all the words in the database have been presented. In our version of the DRT words are played sequentially, one every 2.5 – 3 seconds; the visual presentation precedes the aural presentation by 1sec., and the decision (binary) must be made within 1sec of the aural presentation. Words in the database are divided into "runs", and the duration of one run is limited to about 2.5 minutes (to avoid fatigue).

The scores of one complete DRT-session will be tabulated with a cell granularity of [quadrant×feature], as illustrated in Table A.2. A table-entry contains the number of words per cell that where mistakenly identified; it is an integer between 0 and 4, since the total number of words per cell is 4.

Our knowledge about the acoustic correlates of the Jakobsonian dimensions (Table A.3) provides diagnostic information about temporal representation of speech, while the vowel quadrant identity provides information about the frequency range (i.e. location of the formants in action). Hence, the integrated information can link phonetic confusions with their origin in the time-frequency plane. We shall utilize the usage of such linkage to guide the procedure of tuning the parameters of the auditory model.

A.2 Experiments

We conducted a formal DRT test in quiet and in noise using speech-shape noise at three SPL intensities (70, 60 and 50dB) and at three SNRs (10, 5 and 0dB). We have measured the human performance for the synthetic DRT stimuli and for the naturally spoken DRT stimuli to test the extent to which the usage of synthetic stimuli in the presence of noise worsens human performance[1]. Figure A.1 shows the human performance for naturally spoken stimuli. The figure is structured as a 3×3 matrix with the rows and columns representing SNR and SPL (in dB), respectively. Each panel shows the error distribution, in percent, over the six Jakobsonian dimensions: "VC" is for Voicing, "NS" for Nasality, "ST" for Sustention, "SB" for Sibilation, "GV" for Graveness and "CM" for Compactness. The "+" sign stands for attribute present and the "-" sign for attribute absent. The ordinate is termed "Error", and it represents the percentage of words in the category that, when played to the listener, were judged to be the opposite word in the word pair (i.e., the listener switched to the opposite category). The data is collected from six subjects, all students with normal hearing. Using the same format, figure A.2 shows the human performance for synthetic speech stimuli.

[1] All subjects have zero errors for speech in quiet.

Few observations should be noted. First, the grand-mean errors for synthetic speech stimuli is higher (roughly 4 percentage points). Second, at all SNR levels performance is hardly affected by changes in SPL. As for comparing the error *distribution* along the Jakobsonian dimensions, the error patterns are reasonably similar over the ST, SB, GV and CM dimensions but are markedly different along VC and NS. (This observation is illustrated in Figure A.3 which shows a linear relationship between synthetic speech scores and naturally spoken speech scores for the ST, SB, GV and CM dimensions, but not for VC and NS.) Compared to naturally spoken stimuli, synthetic speech stimuli are easier to be distinguished for Nasality but are harder to be discriminated for Voicing. Furthermore, the asymmetry in errors between the "+" and "-" attributes of VC is in opposite direction. We assume that these differences reflect a failure of the speech synthesizer to generate appropriate acoustic representation of supra-segmental cues (e.g. prosody). We suggest that confusions due to misrepresentation of supra-segmental cues are marginally affected by peripheral processing. Hence, we will proceed with using synthetic speech to tune the PAM.

Table A.1. Samples of word-pairs used in Voiers' DRT (1983).

Voicing (VC) (*Voiced - Unvoiced*)	**Nasality (NS)** (*Nasal - Oral*)	**Sustention (ST)** (*Sustained -Interrupted*)
veal - feel	meat - beat	vee - bee
zed - said	neck - deck	fence - pence
-	-	-
Sibilation (SB) (*Sibilated - Assibilated*)	**Graveness (GV)** (*Grave - Acute*)	**Compactness (CM)** (*Compact - Diffuse*)
cheep - keep	peak - teak	key - tea
jot - got	wad - rod	got - dot
-	-	-

Table A.2. A sample of the outcome of one DRT session, one stimulus condition, and one subject. A table-entry contains the number of words per [quadrant×feature] bin mistakenly identified (an integer between 0 and 4).

	VC		NS		ST		SB		GV		CM	
	+	-	+	-	+	-	+	-	+	-	+	-
High-Front	0	0	1	1	0	4	2	2	2	1	1	1
High-Back	1	1	2	0	2	1	1	0	1	3	0	0
Low-Front	1	0	0	0	1	3	0	1	1	4	1	1
Low-Back	1	1	1	1	3	4	2	3	3	2	1	0

Table A.3. The Jakobsonian dimensions and their acoustical correlates

Voicing	- Periodicity and shorter time of onset duration (Voiced) - Discriminability - at [0,1000] Hz
Nasality	- Formants at 200, 800 and 2200~Hz - Nulls throughout the frequency range (Nasals) - Discriminability - at [0,1000] Hz
Sustention	- Gradual onset and presence of mid-frequency noise (Sustained) - Durational and high-frequency cues
Sibilation	- Higher-frequency noise and greater duration (Sibilant) - Duration is most important acoustical correlate
Graveness	- Origin and direction of second-formant transitions - Grave consonants - steep upward transitions - Acute consonants - downward second-formant transitions - Greater concentration of low-frequency energy (Grave)
Compactness	- Concentration of spectral energy at mid-frequency range (Compact) - More-widely separated spectral peaks (Diffused)

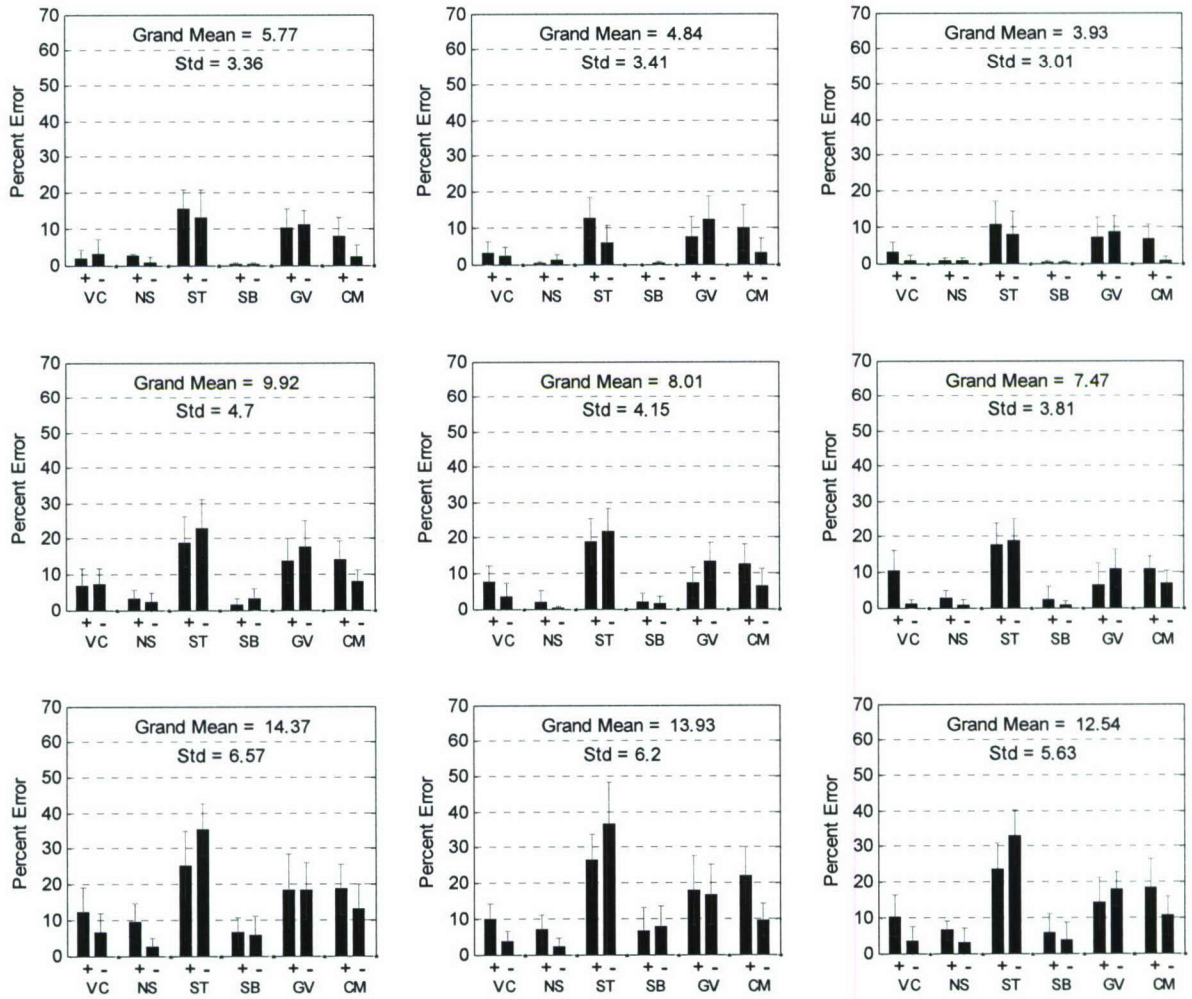

Figure A.1: Human performance on Voiers' 2AFC DRT task using naturally spoken speech. Performance is broken down into DRT dimensions having the attributes of voicing (VC), nasality (NS), sustension (ST), sibilation (SB), graveness (GV), and compactness (CM). + indiciates diphones that have the attribute. – indicates diphones that do not have the attribute. The grand mean is computed by averaging the percent correct over all dimensions and +/- attributes. As SNR decreases, human performance decreases. Human errors moderately decrease as SPL is decreased.

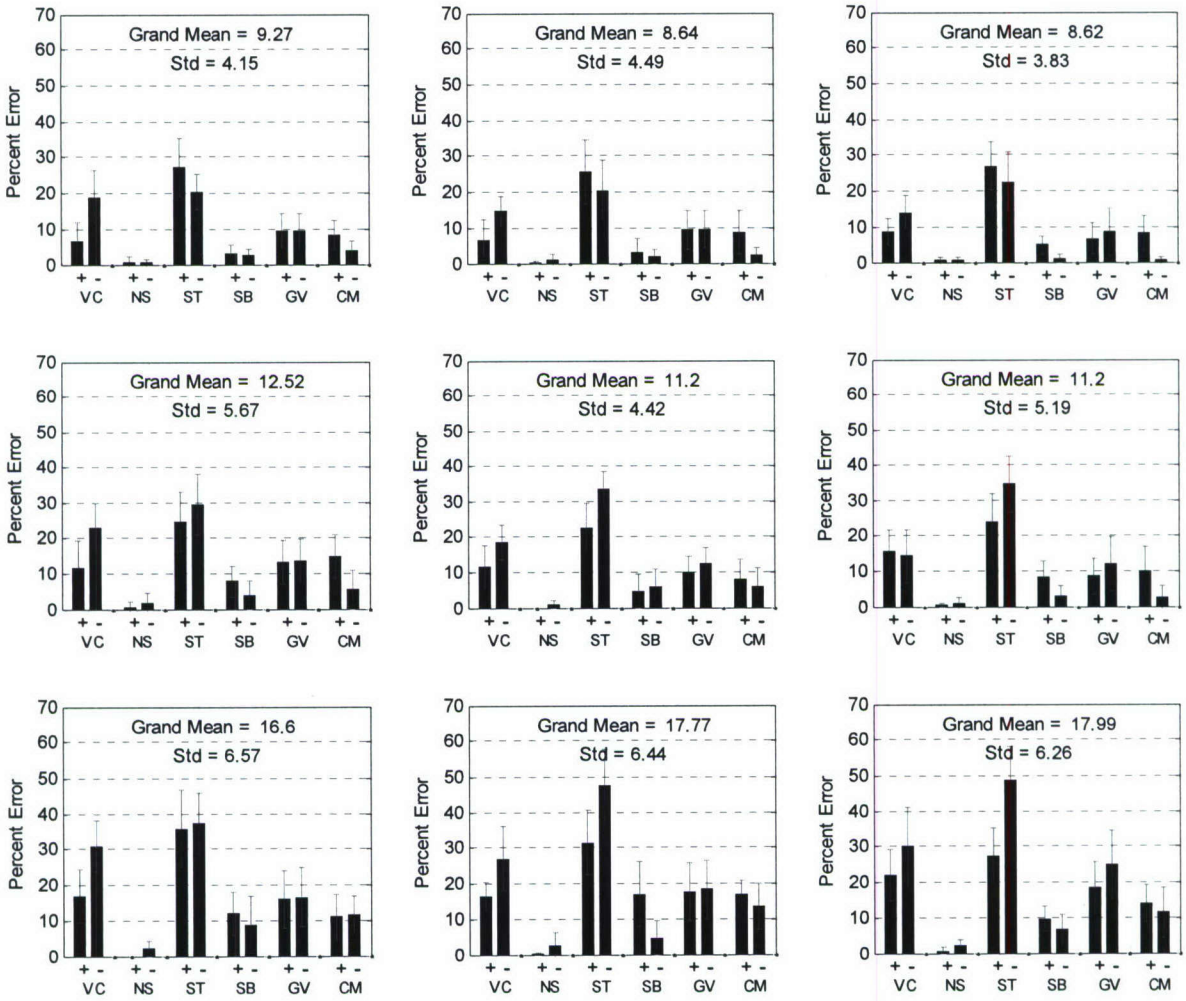

Figure A.2: Human performance on Voiers' 2AFC DRT task using synthetic speech created by the HLsyn speech synthesis system. Performance is broken down into DRT dimensions having the attributes of voicing (VC), nasality (NS), sustension (ST), sibilation (SB), graveness (GV), and compactness (CM). + indiciates diphones that have the attribute. – indicates diphones that do not have the attribute. The grand mean is computed by averaging the percent correct over all dimensions and +/- attributes. As SNR decreases, human performance decreases. Human errors moderately decrease as SPL is decreased for all conditions but the 0dbSNR cases.

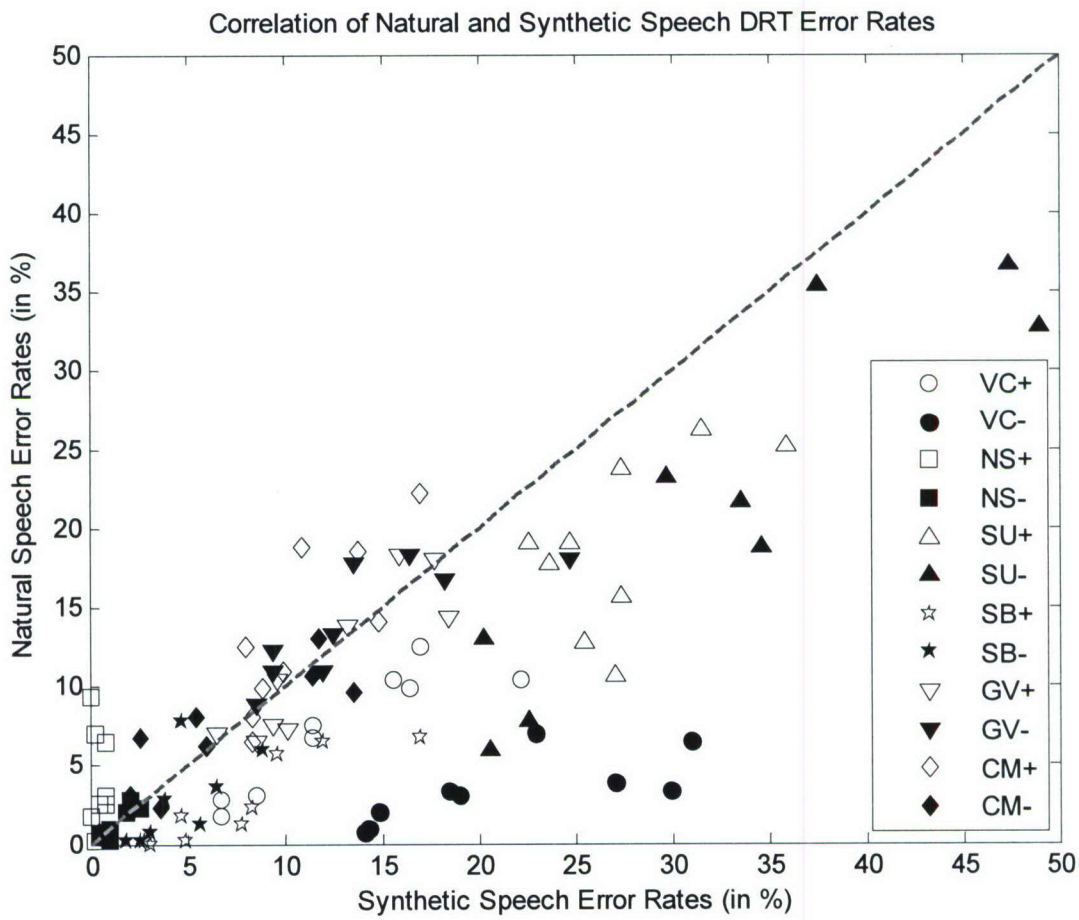

Figure A.3: Correlation of Human and Synthetic Error Rates. Scores are averaged across 6 subjects. Each mark corresponds to the average subject score for a particular Jacobsonian dimension (Voicing, Nasality, Sustension, Sibilation, Graveness, and Compactness) in a particular noise level condition (the conditions examined were noise levels of 70, 60, and 50dBSPL with speech at 10, 5, and 0 dBSNR). The grey dashed line corresponds to an equal error rate between synthetic and natural scores. The voicing and sustension categories in particular are biased towards more synthetic errors. Nasality-plus is biased towards more natural speech errors. The rest of the dimensions (nasality-minus, sibilation, graveness, and compactness) have nearly equal error rates corresponding to a good correlation between synthetic and human error rates.

APPENDIX B. PHONE IDENTIFICATION – HUMAN PERFORMANCE

This experiment examined the consonant confusion patterns of normal-hearing subjects listening to synthetically-produced or naturally-produced consonant-vowel consonants (CVC) under various noise levels, signal-to-noise ratios and filtering conditions.

Speech materials

Naturally-produced speech. The consonant-vowel-consonant (CVC) database is composed of 864 CVC syllables, each preceded by the unstressed schwa vowel /ə/. The materials were recorded by one male and one female talker, each of whom produced half of the CVC tokens. The syllables were constructed using 12 consonants and 6 vowels. The consonants were /p, t, k, b, d, g, θ, v, ð, s, ʃ, z/ and the vowels were /i, ɑ, u, ɪ, ɛ, ʊ/. Out of a total of 864 possible CVC syllables, 840 syllables were selected for the experiment. One half of the syllables were produced by a male talker and the other half by a female talker. The mean durations of the tokens spoken by each of the two talkers were 634 (M) and 574 (F) ms.

Synthetically-produced speech: The database for the synthetically-produced speech was the same database as was used with the naturally-produced speech. The CVC words were synthesized with the help of Ed Bruckert using HLSyn, a modification of the Klatt synthesizer that was developed by Sensimetrics Corporation. As in the naturally spoken corpus, the syllables for the synthetic corpus were constructed using the same 12 initial consonants, 6 vowels and 12 final consonants.

Subjects

A total of 9 subjects started the experiment. Four left the experiment before completion due to time constraints and one subject's data was eliminated because the procedure for collecting the data was changed during the course of the time that she was participating in the study.

Of the 4 people who completed the experiment, 3 were female and 1 was male. All had clinically normal hearing and were native speakers of American English. Two subjects were tested with the naturally-produced speech and two subjects were tested with the synthetically-produced speech.

Conditions

The CVCs were presented in speech-shaped noise at three levels (50 dB, 60 dB and 70 dB SPL) and at three signal-to-noise ratios (-1, 5 and 11 dB) for a total of 9 conditions.

In addition, low-pass (0-2100 Hz) and high-pass (2100-4500 Hz) filtered speech conditions were tested with one noise level (60dB SPL) and one signal-to-noise ratio (5 dB). The low-pass or high-pass filtering was applied to the CVC and the speech-shaped was then added to the filtered signal.

Test Procedures

Two subjects were tested with the naturally-produced stimuli and two subjects were tested with synthetically-produced stimuli. Otherwise all testing was the same for the 4 subjects.

Subjects were seated in front of a computer terminal in a soundproof booth and listened to the speech materials binaurally under Sennheiser HD580 headphones. On the first visit, each subject was given a pure tone hearing test to document normal hearing.

During testing, the listeners knew the constraints of the database (e.g., that whatever token was chosen from the CVC database, it would begin with the unstressed schwa vowel /ə/ and be followed by a Consonant-Vowel-Consonant sequence, where each of the consonants would be one of 12 possible, and each of the vowels would be one of six possible). A copy of the computer graphical user interface (GUI) is shown in Figure 1. Complete testing for each subject consisted of roughly 23 2-hour visits to our lab.

The order of testing was as follows:

(1) Pre-training session(s) - Pre-training consisted of one set of 840 CVCs in quiet without correct-answer feedback. This test served as the baseline (for error patterns) for each subject and to familiarize the subjects with the test procedure.

(2) Training session(s) - Subjects listened to a randomized set of 840 CVCs in quiet with correct-answer feedback.

(3) Post-training sessions – A noise level (50 dB, 60 dB or 70 dB SPL) was chosen at random. Within the chosen noise level, one set of 840 CVCs was presented at each of the 3 SNRs. The same procedure was followed for the remaining noise levels. No correct-answer feedback was provided during these sessions.

(4) Filtered CVCs – For the low-pass and high-pass filtered speech conditions, data was collected with one noise level (60 dB SPL) and one SNR (5 dB). For each of the filtered conditions, subjects listened to a set of 840 CVCs with correct-answer feedback as training before the test condition. The test condition consisted of one set of 840 CVCs

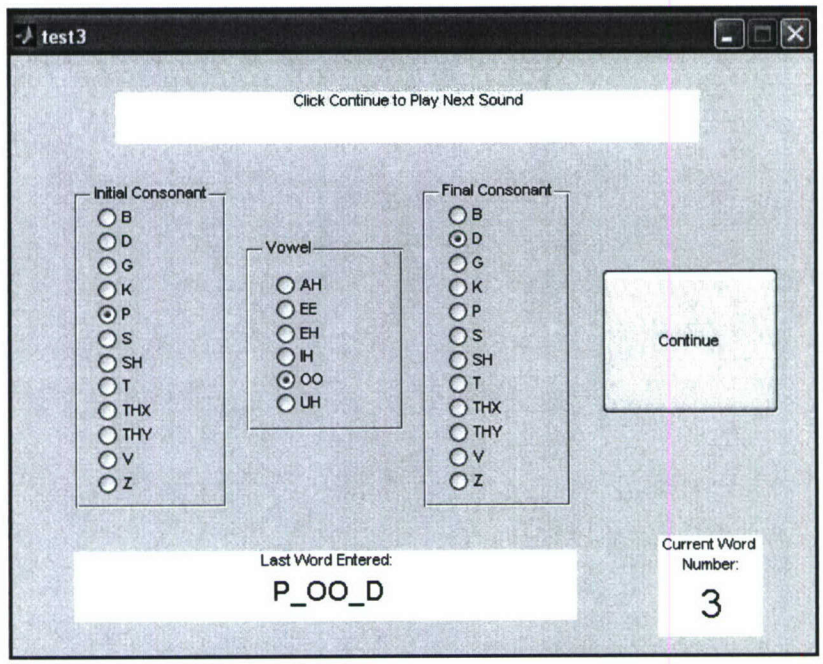

Figure B.1: Graphical User Interface (GUI) used for subject responses in the psychophysical experiment.

REFERENCES

[1] Braida, L. D., (1988). Development of a model for multidimensional identification experiments, *J. Acoust. Soc. Am.* 84, S142.

[2] Braida, L. D., (1991). Crossmodal integration in the identification of consonant segments, *Q. J. Exp. Psychol.* 43A(3) 647-677.

[3] Cohen, M.M. and Massaro, D.W., (1995). Perceiving visual and auditory information in consonant-vowel and vowel syllables, in Sorin, C., Meloni., H., and Schoentgen, J. (Eds.), *Levels in Speech Communication: Relations and Interactions: A Tribute to Max Wajskop*, Amsterdam: Elsevier Science B. V.

[4] Dewson, J. H. 1968. Efferent olivocochlear bundle: Some relationships to stimulus discrimination in noise. *J. Neurophysiol.* 31, 122–130.

[5] Ghitza, O. (1993). Processing of spoken CVCs in the auditory periphery: I. Psychophysics. *JASA*, 94(5), 2507-2516.

[6] Giraud, A. L., Garnier, S., Micheyl, C., Lina, G., Chays, A., Chery-Croze, S. (1997). Olivocochlear efferents involved in speech-in-noise intelligibility. *Neuroreport* 8, 1779-1783.

[7] Goldstein, J.L. (1990). Modeling rapid waveform compression on the basilar membrane as a multiple-bandpass-nonlinearity filtering. *Hear. Res.* 49, 39-60.

[8] Greenberg, S. (1999). Speaking in shorthand - A syllable-centric perspective for understanding pronunciation variation. *Speech Communication* 29, 159-176.

[9] Greenberg, S. (ed.) (1988). *Representation of Speech in the Auditory Periphery. J. Phon.* 16, 1-149

[10] Guinan, J. J. (1996). Physiology of olivocochlear efferents. In: Dallos, P., Popper, A. N. Fay, R.R., (eds). *The Cochlea*, New York: Springer Verlag, 435-502.

[11] Hant, J.J., and Alwan, A. (2003). A psychoacoustic-masking model to predict the perception of speech-like stimuli in noise. *Speech Communication*, 40, 291-313.

[12] Hopfield, J.J. (2004). Encoding for computation: Recognizing brief dynamical patterns by exploiting effects of weak rhythms on action-potential timing. *Proc. Nat. Acad. Sci.* 101, 6255-6260.

[13] Jakobson, R., Fant, C. G. M., and Halle, M. (1952). Preliminaries to speech analysis: the distinctive features and their correlates. Technical report, Acoustic Laboratory, Massachusetts Institute of Technology.

[14] Lippmann, R.P., Braida, L.D., and Durlach, N.I. (1981). A Study of Multiband Amplitude Compression and Linear Amplification for Persons with Sensorineural Hearing Loss, *J. Acoust. Soc. Am.* 69, 524-534.

[15] May, B.J., Sachs, M.B. (1992). Dynamic range of neural rate responses in the ventral cochlear nucleus of awake cats. *J. Neurophysiol.* 68, 1589–1603.

[16] Ronan, D, Dix, A, Shah, P. and Braida, L.D. (2004). Integration of acoustic cues for consonant Identification across frequency bands, *J. Acoust. Soc. Am.* 116, 1749-1762.

[17] Voiers, W. D. (1983). Evaluating processed speech using the diagnostic rhyme test. *Speech Techn.* 1 30–39.

[18] Winslow, R.L., Sachs, M.B. (1988). Single-tone intensity discrimination based on auditory-nerve rate responses in backgrounds of quiet, noise, and with stimulation of the crossed olivocochlear bundle. *Hearing Res.* 35,165–190.

[19] Zeng, P. G., Martino, K. M., Linthcum, F. H., Soli, S. (2000). Auditory perception in vestibular neurectomy subjects. *Hearing Res.* 142, 102-112.